事例の論理構造に基づく関係的知識発見に関する研究

和田・犬塚 研究室　元山 純一

No. 17417598　Motoyama Jun-ichi

1 はじめに

データマイニングとは，大量のデータから隠された知識や新しい規則を発見するプロセスである．

データマイニングの中でも，複数の関係表で表されたデータベースから知識や規則を取り出すプロセスを関係データマイニングという．一般的なデータはデータ間に関係性を持っており，複数の関係表をもつデータベースで表される．よって，実際のデータから隠された知識や新しい規則を発見するためには，関係データマイニングをおこなう必要がある．

関係データマイニングをおこなう手法として，**帰納論理プログラミング** (Inductive Logic Programming : ILP) が注目される．これは論理的な記述によって高い可読性を持ち，関係データマイニングの有力な手法と考えられている．

ILP の枠組みにおけるデータマイニング (以下, ILP データマイニング) では，多くの手法が提案されてきた．その中の一つに WARMR がある [1]．WARMR は，頻出なパターンをすべて枚挙することができるが，生成される候補は多大にあり，処理に多くの時間がかかってしまうという問題点がある．

本研究では，高速に処理をおこなう ILP データマイニング手法を提案する．これは対象となる事例に見られる性質に注目し，**(1) 事例から性質を抽出し，(2) 性質の組み合わせたパターンを生成し頻出なものを枚挙する手法**である．

2 手法のアイディア

以下の家族関係のデータベース $\mathbf{R}_{\mathrm{fam}}$ があるとする．

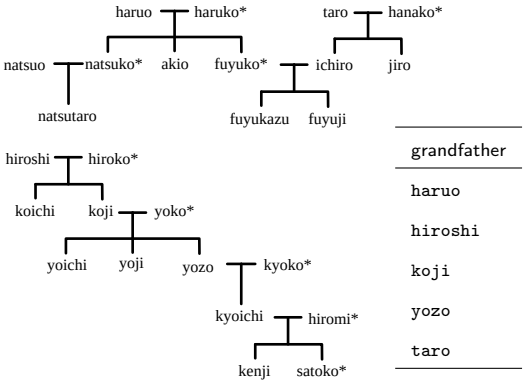

図 1: 家族関係のデータベース $\mathbf{R}_{\mathrm{fam}}$

これは，親子関係を表す parent と性別を表す male, female の複数の表を持つ関係データベースで表されており，x は y の親であるという parent(x, y) と x の性別を表す male(x)，female(x) という述語で構成されている．*がついている名前は female を満足するものである．また目標事例として grandfather の表も与えられている．目標事例はその特徴が未知であるか，その特徴が分かっていても，実際の例に見られる特徴に興味があるとする．このような関係データベースから，事例に見られる特徴をパターンとして取り出すことが目的である．

例えば，この家族関係 $\mathbf{R}_{\mathrm{fam}}$ のデータベースにおいて事例 grandfather(koji) に見られる述語の組として以下のようなものがある．

$$\mathsf{grandfather}(\mathtt{koji}) \leftarrow \mathsf{parent}(\mathtt{koji}, \mathtt{yozo}) \wedge \mathsf{parent}(\mathtt{yozo}, \mathtt{kyoichi}) \wedge \mathsf{male}(\mathtt{kyoichi}).$$

これは「koji が kyoichi という孫息子をもつ」という事実である．このように，事例が表す述語の組の中でも意味のある述語の組を**性質**と考える．この性質を使用して，意味のあるパターンに限定してマイニングを行うことが本提案手法のアイディアである．

3 提案手法

3.1 性質

性質を定義するために，モードを導入する．モードとは，述語の各引数が入力引数であるのか出力引数であるのかを表す情報で，入力引数を表す**入力モード** $\oplus$ と出力引数を表す**出力モード** $\ominus$ がある．例えば，先ほどの家族関係 $\mathbf{R}_{\mathrm{fam}}$ に現れる述語のモードは，parent($\oplus$, $\ominus$)，male($\oplus$)，female($\ominus$) であると仮定する．

モードについて注目すると，述語は 2 つのクラスに分けられる [2]．第 1 のクラスは，すべての引数が入力モードで**判定述語**という．この述語はすべての引数に値の定まった形で呼ばれ，その真偽を決定する．

もう一つのクラスの述語は，入力モードと出力モードを持つ述語のクラスで**経路述語**という．これは，入力モードである引数の値が束縛されて呼び出され，出力モードの引数に値を返す関数的な使い方をする．

例えば，家族関係 $\mathbf{R}_{\mathrm{fam}}$ では male，female は判定述語，parent は経路述語である．また，各述語で構成されたリテラルをそれぞれ**判定リテラル**，**経路リテラル**と呼ぶ．

判定述語，経路述語の考え方を使って，先ほどの述語の組に注目すると，次のことがいえる．まず，(1) 事実を表す「判定リテラル」が一つだけある．(2)「経路リテラル」によって，引数が事例から判定リテラルまで鎖上に繋がっていることで，事例についてのある事実を表している．

このような性質をいくつかの事例から取り出し，得られた性質を使って興味深いパターンを取り出すことを考える．

3.2 MAPIX

MAPIX では，それぞれの性質について，その性質を持つかどうかの真偽を一つの属性と考えることで，相関ルールマイニングをおこなう．相関ルールマイニングとは，マーケットバスケット分析を目的として提起された方法論で，マーケットで売られている個々の商品について，どのように買われているかの規則性を見つけることが目的である．相関ルールは，効率的な探索をおこなう手法として Apriori アルゴリズム [3] が考案されている．

MAPIX では，**(1) 性質をいくつかの事例から取り出し，(2)Apriori アルゴリズムを利用して，性質を組み合わせて頻出なパターンを枚挙する．**

これによって，より高速に ILP の枠組みでマイニングをおこなうことができるが，問題点として得ることのできないパターンが存在する．

3.3 DUPLIPIX

DUPLIPIX は，MAPIX の出力するパターンの問題点を解決する手法である．まず，その問題点について具体的な例を出して述べる．

例 1 以下のような事例について考える．

$$\mathsf{grandfather}(\mathtt{koji}) \leftarrow \mathsf{parent}(\mathtt{koji}, \mathtt{yozo}) \wedge \mathsf{male}(\mathtt{yozo}) \wedge \mathsf{parent}(\mathtt{yozo}, \mathtt{kyoichi}) \wedge \mathsf{male}(\mathtt{kyoichi}).$$

この事例は，「息子をもつ」という性質「$\mathsf{grandfather}(A) \leftarrow \mathsf{parent}(A,B) \wedge \mathsf{male}(B)$」と「孫息子をもつ」という性質「$\mathsf{grandfather}(A) \leftarrow \mathsf{parent}(A,B) \wedge \mathsf{parent}(B,C) \wedge \mathsf{male}(C)$」を満たす．それぞれの性質を p_1，p_2 としたとき，事例 koji は p_1，p_2 をどちらも満たしているので，$p_1 \wedge p_2$ を満たすと考えることができる．しかし，$p_1 \wedge p_2$ は「息子をもつ ∧ 孫息子をもつ」という意味であり，事例 koji に現れる「息子と孫息子がおり，孫息子は息子の息子である」という意味を成さない．

よって DupliPix では，性質を使って事例の部分的な構造を生成することを考える．具体的には，「調べたい性質の組み合わせを実際に事例に見られるような述語の組で取り出す方法」を提案し，それらの複数の性質をパターンと考えマイニングをおこなう．

このアルゴリズムの問題点として，論理的に同値なパターンが出力されるため，それらを出さないように削除するなどして制御する必要はある．この処理には，多くの時間がかかってしまうことである．

3.4 EQUIVPIX

前節の DupliPix アルゴリズムによって，事例に見られるすべての構造をマイニングすることができる．しかし，ILP データマイニングでは事例には見られない構造でも，事例が満たすならば出力したいという要求がある．例えば，以下のような場合がある．

例 2 家族関係 $\mathbf{R}_{\mathrm{fam}}$ で事例 yozo について考える．

$$\mathsf{grandfather}(\mathtt{yozo}) \leftarrow \mathsf{parent}(\mathtt{yozo}, \mathtt{kyoichi}) \wedge \mathsf{male}(\mathtt{yozo}) \wedge \mathsf{parent}(\mathtt{kyoichi}, \mathtt{kenji}) \wedge \mathsf{male}(\mathtt{kyoichi}) \wedge \mathsf{male}(\mathtt{kenji})$$
$$\mathsf{parent}(\mathtt{kyoichi}, \mathtt{satoko}) \wedge \mathsf{female}(\mathtt{satoko}).$$

この事例からは以下のようなパターンは，事例の本来の構造に無いため出力されない．

$$p = \mathsf{grandfather}(A) \leftarrow \mathsf{parent}(A,\ B) \wedge \mathsf{male}(B) \wedge \mathsf{parent}(A,\ C) \wedge \mathsf{parent}(C,\ D) \wedge \mathsf{male}(D).$$

しかし，$A = \mathtt{yozo}$，$B = \mathtt{kyoichi}$，$C = \mathtt{kyoichi}$，$D = \mathtt{kenji}$ とすることで，事例 yozo はパターン p を満たすことがわかる．

DupliPix では出力できなかったこのようなパターンは，例 1 の「息子がいる ∧ 孫息子がいる」というような性質を連言で表したようなパターンである．このような要求を満たすために，EquivPix では，**事例の部分的な論理構造を連言にもつパターン**を効率よくマイニングすることを考える．また，DupliPix の問題点を解決するために，**同値な性質を一つのグループにし，グループ単位で組み合わせていく**ことで，同値なパターンを出さないように処理をおこなう．

同値なパターンの排除

これら三つの手法は，論理的に同値なパターンは意味的にも重複していると考え，出力しないように構成し証明を与えた．これによって，無駄な処理を省いている点で従来法より優れている．

4 実験

本実験では，従来手法 Warmr と提案した MaPix，EquivPix との処理にかかる時間と得られるパターンの比較をおこなう．DupliPix は処理に多くの時間がかかってしまうため，ここでは比較の対象から除外する．実験データとして，ILP における代表的なベンチマークである Bongard を使用した．

実験結果として，処理にかかった時間と得られたパターンの数，さらにそのパターンを同値類に分けたときの同値類の数を表 1 に示す．

表 1：実行時間とパターン数と同値なパターン

閾値=5%	時間 (s)	パターン数	同値類の数
Warmr	1098.5	5480	782
MaPix	142.6	160	160
EquivPix	237.7	625	625

表 1 から，提案手法は Warmr より少ない時間で処理していることがわかる．また Warmr は同値なパターンが多く存在し，このことによって処理に時間がかかると考えられる．一方，提案手法では，同値なパターンを出さずに処理をおこなうため，効率的に探索をしていることがわかる．

しかし，提案手法は性質という述語の組に制限を与えているため，Warmr と比較して出力していないパターンも存在することがわかる．

5 おわりに

本研究では，データマイニングに注目し，帰納推論と論理プログラミングを結合した強力なアプローチである ILP を使用することで，データマイニングに重要な可読性の高い知識を得る三つの手法を提案した．提案手法では性質というものに注目し，性質を使うことで従来手法より高速にマイニングをおこなうことを可能にした．

今後の課題として，従来手法では得られるパターンのすべてを出力していないため，それらを取り出すように改善する必要がある．

参考文献

[1] L. Dehaspe, H. Toivonen :“Discovery of Relational Association Rules”, in Relational Data Mining, pp. 189–212, Springer, 2001.

[2] M. Furusawa, N. Inuzuka, H. Seki, H, Itoh : “Induction of Logic Programs with More Than One Recursive Clause by Analysing Saturations”, ILP-97, pp. 165–172, LNAI 1297, Springer, 1997.

[3] R. Agrawal, R. Srikant : “Fast Algorithms for Mining Association Rules”, Proc. VLDB, pp. 487-499, Morgan Kaufmann, 1994.

発表論文

- 元山 純一，中野 智文，犬塚 信博 : “関係的相関ルール導出のための事例の性質の抽出”, FIT2005，一般論文集.
- 元山 純一，中野 智文，犬塚 信博 : “関係的相関ルール導出のための事例の性質の抽出”, 第二回 情報科学ショップ，2006.
- 浦澤 真平，元山 純一，中野 智文，犬塚 信博 : “関係的知識発見手法を用いたマイニング応用”, 第二回情報科学ショップ，2006.
- Jun-ichi Motoyama, Shinpei Urazawa, Tomofumi Nakano, Nobuhiro Inuzuka : “A mining algorithm using property items extracted from sampled examples”, The 16th International Conference on Inductive Logic Programming (ILP 2006), Santiago, Spain, August 24-27, 2006.